ADVANCE STATISTICAL METHODS AND TESTING OF HYPOTHESIS

Dataset: SEPHORA WEBSITE DATA

ANJALI CHAUHAN
3154
ADVANCE STATISTICS & HYPOTHESIS TESTING
2022 - 2023
Prof: SABITHA PRAVEEN

1.1 INTRODUCTION

Sephora is a French multinational retailer of personal care and beauty products. A leader in prestige omni-retail, our mission at Sephora is to create a welcoming beauty shopping experience for all and inspire fearlessness in our community.
We operate over 2,700 stores in 35 countries worldwide, with an expanding base of over 500 stores across the Americas, and a world-class e-commerce site.

Sephora offers beauty products including cosmetics, skincare, body, fragrance, nail color, beauty tools, body lotions, and haircare. Since opening our first US store in New York's SoHo neighborhood in 1998, Sephora has been an industry-leading champion of diversity, inclusivity, and empowerment in the US, guided by our longstanding company values.

The company was founded in Limoges in 1969 and is currently based in Neuilly-sur-Seine, France. Sephora is owned by luxury conglomerate LVMH as of 1997. The name comes from the Greek word meaning beauty, Sophos.

Sephora was first launched in Paris in August 1970. It was acquired by Dominique Mandonnaud in 1993, who merged the purchase with his own perfume chain under the Sephora brand. Mandonnaud is credited for designing and executing Sephora's "assisted self-service" sales experience, which separated itself from standard retail models for cosmetics by encouraging customers to test products in retail locations before purchasing.
In 1999 Sephora launched its first website in the US, taking accessibility to new heights.
2016 Seophra become the #1 prestige beauty retailer in the US. In 2018 Sephora was awarded as the best retainer by the world retailer Congress.2021 Sephora joins the Interbrand ranking of the 10 most wannabe brands worldwide.

We believe that beauty thrives in diversity and discovery.
Our purpose is to expand the way the world sees beauty by empowering the ExtraOrdinary in each of us.

Website link: **https://www.sephora.com/**

2.1 OBJECTIVES

- Is there a significant difference between the price and valued price of products on the Sephora website.

- Does the Number of reviews received on the product make a difference in the rating.

- To verify if there is an association between online sales & exclusive sales of products by brand on the Sephora website.

- To inspect whether Brands on Sephora release many Limited edition products for sale.

- To authenticate whether the marketing flag used for marketing affects the brand name on the Sephora website.

3.1 RESEARCH METHODOLOGY

3.1.1 Method of Data Collection

Secondary Data

Data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to estimate the outcome of the situation.

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

3.1.2 Statistical Description

- Mean

Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers.

- Median

The median is the middle number in a sorted, ascending, or descending list of numbers and can be more descriptive of that data set than the average. It is the point above and below which half (50%) the observed data falls, and so represents the midpoint of the data.

- Max

The maximum value in a set of values, excluding any outliers. Both are computed using either the interquartile rule or a user-defined statistical limit.

- Min

the minimum value in a set of values, excluding any outliers. Both are computed using either the interquartile rule or a user-defined statistical limit.

- Standard deviation

Standard Deviation is a measure that shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a “typical” deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set.

- Cental tendency

A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution.

- 1st inter Quartile Range

The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order.

- 3rd inter Quartile Range

The lower quartile, or first quartile (Q1), is the value under which 25% of data points are found when they are arranged in increasing order. The upper quartile, or third quartile (Q3), is the value under which 75% of data points are found when arranged in increasing order.

3.1.3 Hypothesis Used

- Normality Test

Normality tests are used to determine if a data set is well-modeled by a normal distribution and to compute how likely it is for a random variable underlying the data set to be normally distributed.

- Wilcoxon Rank test

Wilcoxon rank-sum test is used to compare two independent samples, while Wilcoxon signed-rank test is used to compare two related samples, matched samples, or to conduct a paired difference test of repeated measurements on a single sample to assess whether their population means ranks differ

- ANOVA Test

ANOVA is to test for differences among the means of the population by examining the amount of variation within each sample, relative to the amount of variation between the samples. Analyzing variance tests the hypothesis that the means of two or more populations are equal.

- Chi-Square Test

A Chi-square test is a hypothesis testing method. Two common Chi-square tests involve checking if observed frequencies in one or more categories match expected frequencies.

- Correlation Test

Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it’s a measure of how things are related. The study of how variables are correlated is called correlation analysis.

4.1 DATA ANALYSIS

4.1.1 Statistical Descriptive analysis

- **Sephora Website Data set:**

Column name	**datatype**	**Description**
Id	Int	Products id on Sephora website
Brand	Object	The brand of the product at Sephora's website
Category	Object	Category of the product at Sephora's website
Name	Object	The name of the product
Rating	Float	The rating of the product
Number of reviews	int	The number of reviews of the product
Love	int	The number of people loving the product
Price	Float	The price of the product with discount and offer
Valued price	Float	The value price of product
Marketing Flags	Boolean	The Marketing Flags of the product from the website if they were exclusive or sold online only
Marketing Flags content	Object	The kinds of Marketing Flags of the product
details	Object	The details of the product available on the website
Online	Int	If the product is sold online only
Exclusive	Int	If the product is sold exclusively on Sephora's website
Limited Edition	Int	If the product is limited edition

Limited time offer	Int	If the product has a limited time offer

- **Dimension**

Rows	**Cols**
1499	16

- **Column name**

Column name
Id
Brand
Category
Name
Rating
Number of reviews
Love
Price
Valued price
Marketing Flags
Marketing Flags content
details
Online
Exclusive
Limited Edition
Limited time offer

- **Summary**

Col name	**Min**	**1st Qu**	**Median**	**mean**	**3rd Qu**	**Max**
rating	0.00	4.00	4.50	4.05	4.50	5.00
No of reviews	0.0	13.0	58.0	337.9	228.5	19000.0
Love	0	1900	5100	17442	14650	660000
price	5.00	26.00	36.00	53.55	64.00	370.00
value_price	5.00	27.00	37.00	55.00	65.00	483.00
Online_only	0.0000	0.0000	0.0000	0.2562	1.0000	1.000
exclusive	0.0000	0.0000	0.0000	0.1594	1.0000	1.000
Limited_edition	0.0000	0.0000	0.0000	0.08672	1.0000	1.000
Limited _time_offer	0	0	0	0	0	0

4.2 Hypothesis Testing

SHAPIRO TEST

To check if the data follows normal distribution or not.
H0: Data follows normal Distribution
H1: Data does not follow normal Distribution
pro_price= data$price
shapiro.test(pro_price)

data	test	p-value
Price	0.72693	2.2e-16

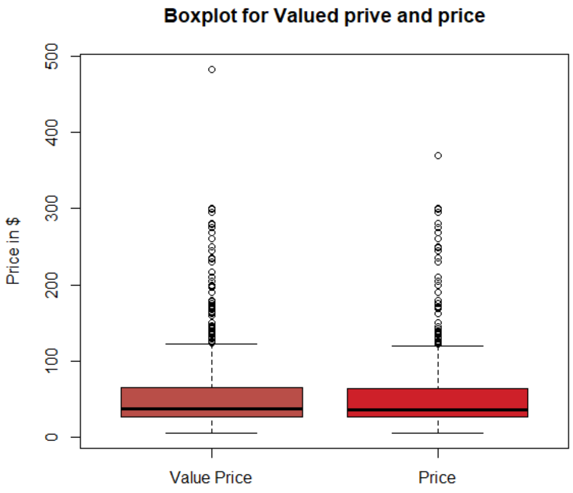
p value is smaller than 0.05, Hence we reject H0.
which implies that the data doesn't follow a normal distribution.

4.2.1 Obj (1) Is there a significant difference between the price and valued price of products on the Sephora website

H0: There is no significant difference between price and value_price
H1: There is a significant difference between price and value_price
wilcox.test(data$value_price, data$price)

data	test	p-value
Price	1145639	0.3501

p-value is greater than 0.05 (0.3501> 0.05), Hence we accept H0.
This supports that there is a significant difference between price and value_price



The above boxplot significantly shows that the actual value of the price is more than the price at which the product is sold on the website.

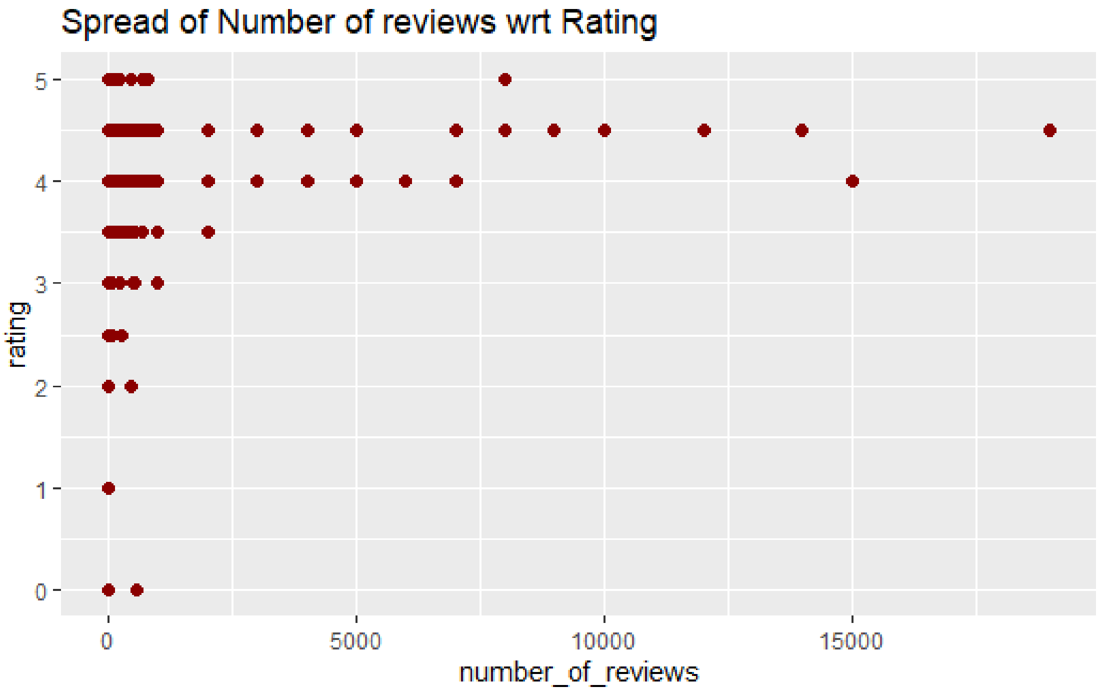
4.2.2 Obj (2) Does the Number of reviews received on the product make a difference in rating

H0: Number of reviews does not make a difference on rating
H1: Number of reviews makes a difference on rating
wilcox.test(data$number_of_reviews, data$rating)

data	test	p-value
(data$number_of_reviews, data$rating)	0.72693	2.2e-16

p value is smaller than 0.05, Hence we reject H0.

This confirms that the number of reviews makes a difference



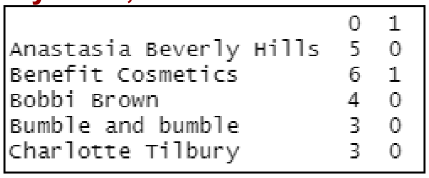
This clarifies that the product with a high number of reviews has a high rating as well.
Clearly, most products have ratings of 4 and 5 and very few have ratings of 3
Products have more than 15000 reviews with a rating in the range between 4 and 5

4.2.3 Obj (3) To verify if there is an association between online sales & exclusive sales on products by brand on Sephora website

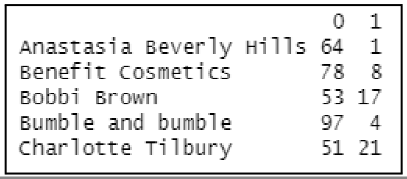
H0: There is no association between online sales & exclusive sales
H1: There is an association between online sales & exclusive sales
offer=table(data$online_only, data$exclusive)
offer
chisq.test(offer)

data	x-squared	df	p-value
(data$online_only, data$exclusive)	1.9886	1	0.1585

p-value = 0.1585
p-value is greater than 0.05(0.1585> 0.05), Hence we accept H0.
this conveys that, there is no association between online sales & exclusive sales



```
                        0  1
Anastasia Beverly Hills 5  0
Benefit Cosmetics       6  1
Bobbi Brown             4  0
Bumble and bumble       3  0
Charlotte Tilbury       3  0
```



```
                         0  1
Anastasia Beverly Hills 64  1
Benefit Cosmetics       78  8
Bobbi Brown             53 17
Bumble and bumble       97  4
Charlotte Tilbury       51 21
```

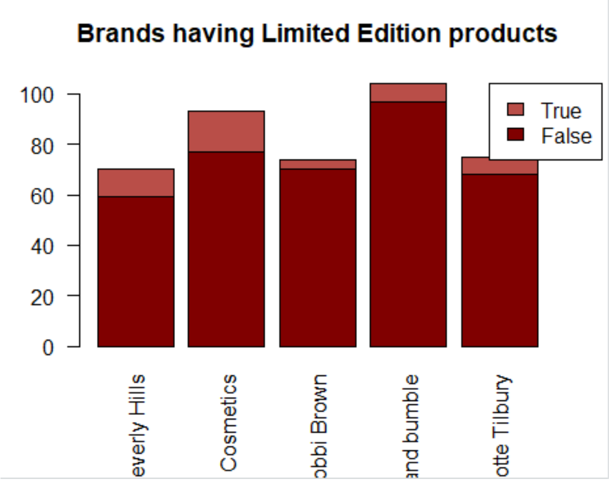
4.2.4 Obj (4) To inspect whether Brands on Sephora release many Limited edition products for sale**.**

#H0: Brand releases limited edition products on Sephora.
#H1: Brand releases less limited edition products on Sephora.

```
data$brand <- as.factor(data$brand)
a=table(data$online_only,data$exclusive, data$limited_edition, data$limited_time_offer)
res=aov(data$limited_edition ~ data$brand)
summary(res)
```

data	df	Sum Sq	Mean Sq	F value	p-value
data$limited_edition ~ data$brand	53	6.78	0.12786	1.65	0.00255
Residuals	1445	111.95	0.07747		

p-value = 0.00255
p-value is smaller than 0.05, Hence we reject H0.
This indicates that limited edition products are launches less on Sephora by brands.



These are the top 5 brands with the most product on the Sephora website and most of them doesn't often produce limited edition products

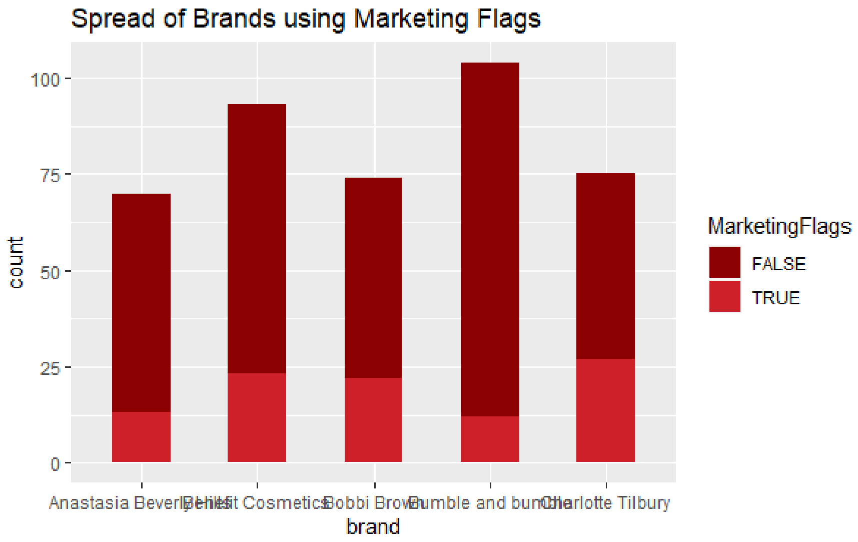
4.2.5 Obj (5) To authenticate whether the marketing flag used for marketing affects the brand name on the Sephora website

H0: There is no association between brand & marketing flags
H1: There is an association between brand & marketing flags

```
market=table(data$brand, data$MarketingFlags)
market
chisq.test(market)
```

data	x-squared	df	p-value
(data$brand, data$MarketingFlags)	513.34	53	2.2e-16

p value is smaller than 0.05, Hence we reject H0.
this implies that there is an association between brand & marketing flags.



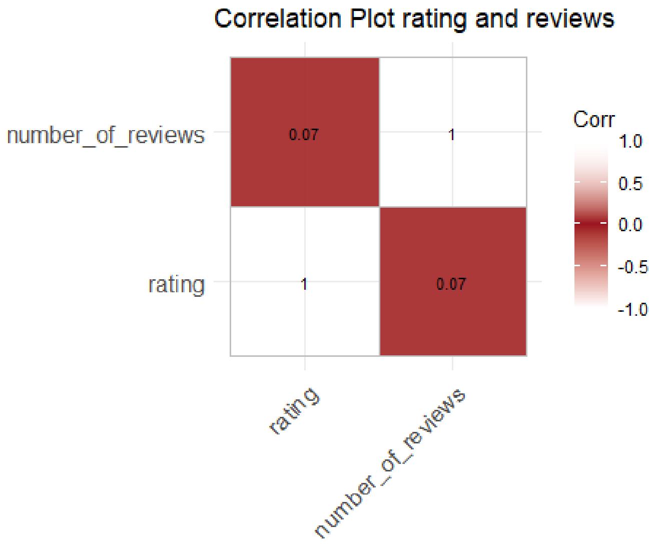
This plot indicates that most brand doesn’t use any marketing flags in the sales of the product.
The brand name makes the product run in the market.

Correlation test

H0: No of reviews do not affect Rating
H1: No of reviews affect Rating
cor.test(data$rating, data$number_of_reviews)

data	df	T-test	p-value	correlation
(data$rating, data$number_of_reviews)	1497	2.8219	0.004836	0.07274199

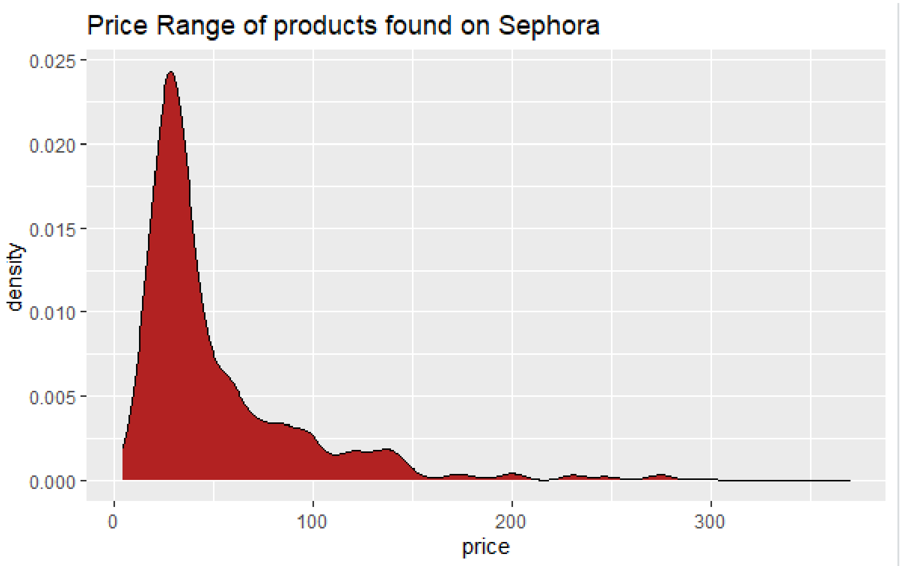
p-value is smaller than 0.05 (0.07274199 > 0.05), Hence we reject H0.
This shows there is very less correlation between the number of reviews affecting the rating(7%).



The plot shows that there is a positive correlation between ratings and the number of reviews.
This means that if the number of reviews increases the rating of the product also increases to some extent.

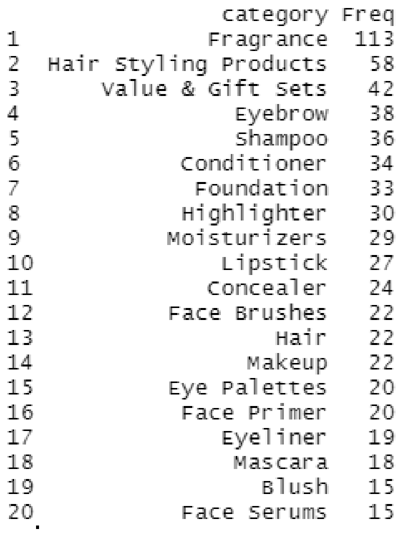
4.3 EDA

4.3.1 Price range on products on the Sephora website



The Range of products found on Sephora is in the range of 0 to 100 Dollars. A few products exceed 100 dollars.

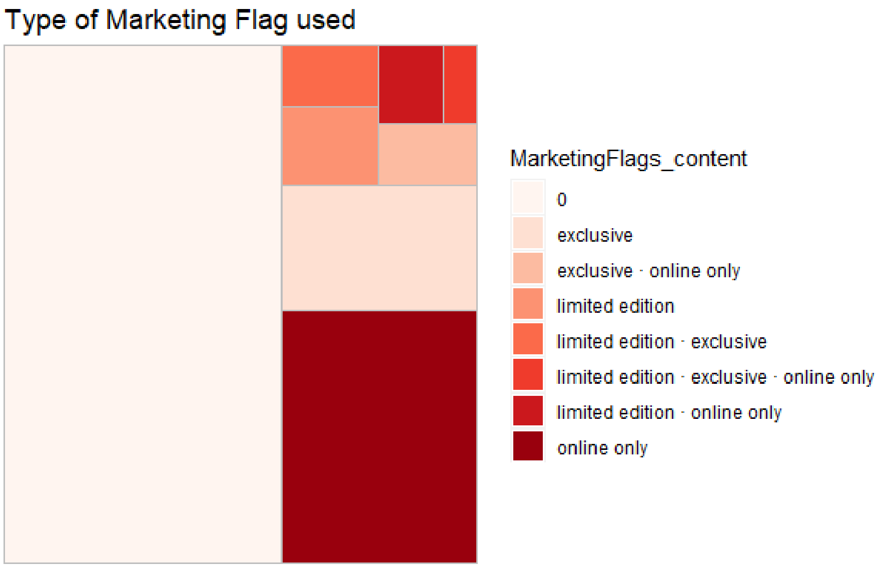
4.3.2 Top 20 Categories of products on the Sephora website



```
                 category Freq
1               Fragrance  113
2    Hair Styling Products  58
3       Value & Gift Sets   42
4                 Eyebrow   38
5                 Shampoo   36
6             Conditioner   34
7              Foundation   33
8             Highlighter   30
9            Moisturizers   29
10               Lipstick   27
11              Concealer   24
12           Face Brushes   22
13                   Hair   22
14                 Makeup   22
15           Eye Palettes   20
16            Face Primer   20
17               Eyeliner   19
18                Mascara   18
19                  Blush   15
20           Face Serums    15
```

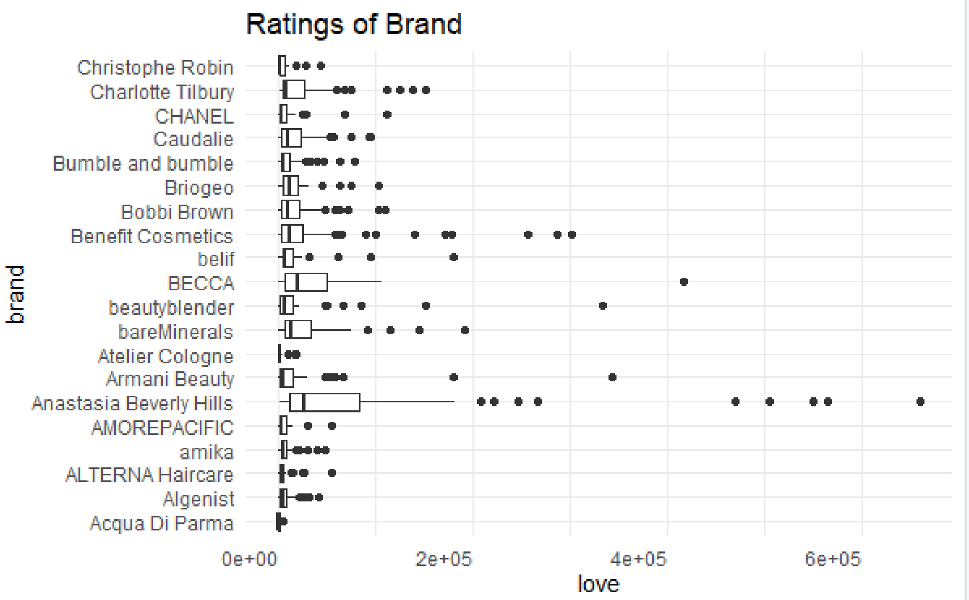
Most products on the Sephora website are f fragrance followed by Hair styling products and value and gift sets.

4.3.3 Types of Marketing flags used for the product by prestige brands



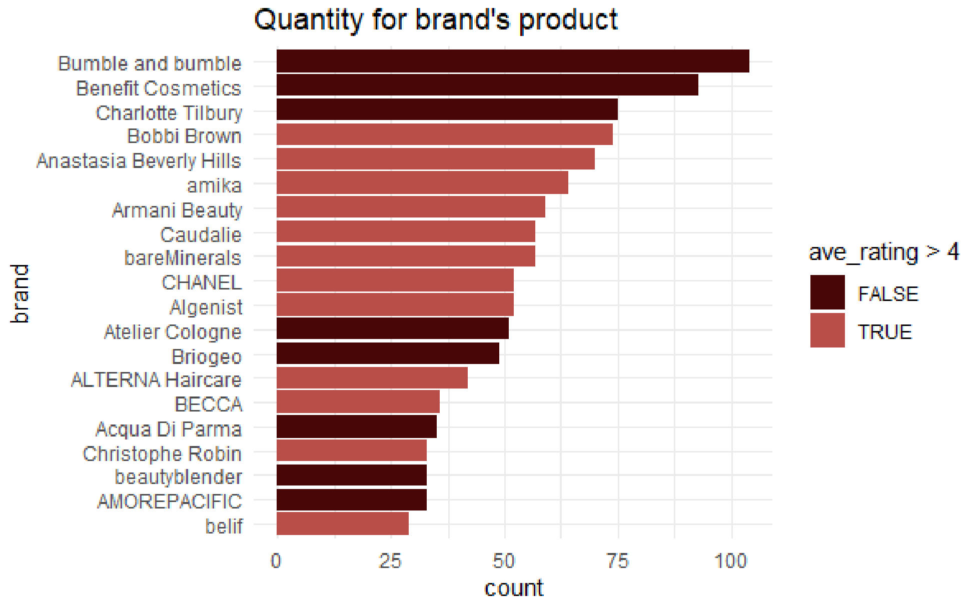
Most brands have products selling on Sephora without and marketing strategy running for it. Second most marketing of products is done by only selling them in the online store. Then exclusive which are only sold on the Sephora website and not by any other retailers

4.3.4 Top 20 brands according to the love(likes) received



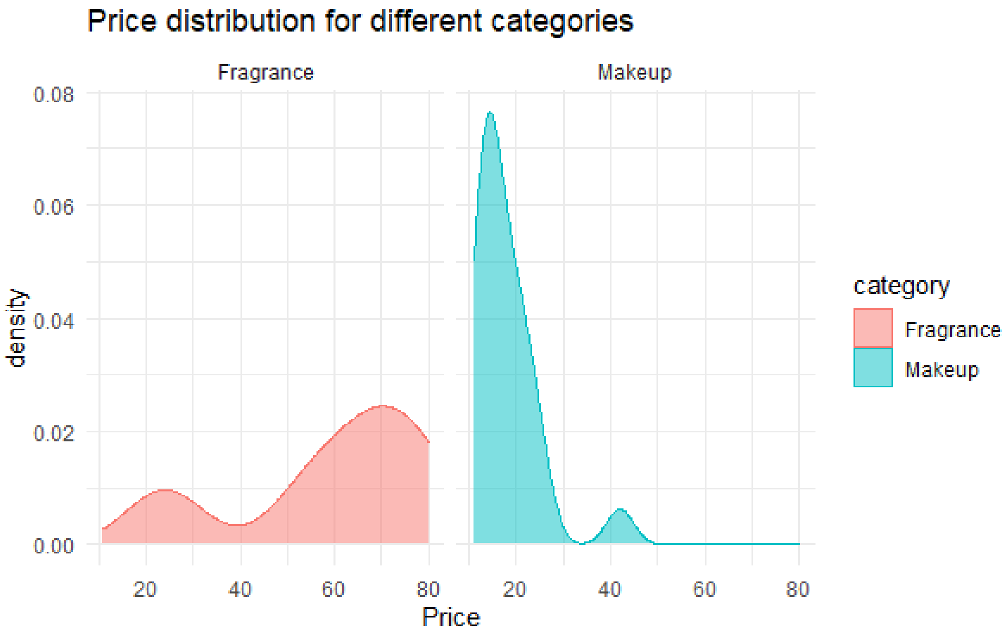
Anastasia Beverly hills product has the most love on the Sephora website. BECCA and BareMinerals brands are second and third top 20 products with high love recorded.

4.3.5 Top 20 brands according to most products sold online only



The Bobbi Brown brand has the most products of them selling online. Coming up next is Anastasia Beverly hills and Amika as the second and third brands of most products online. Bumble and bumble, Benefit Cosmetics, and Charlotte Tilbury brands selling on the Sephora store and not particularly online.

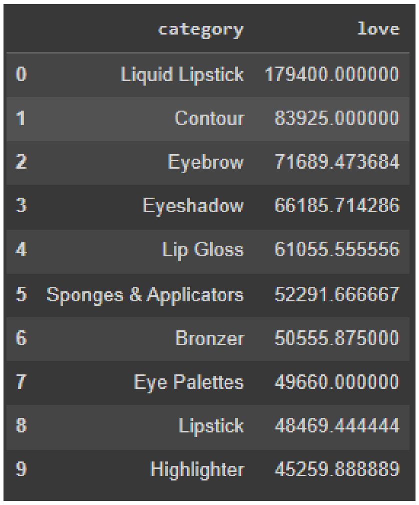
4.3.6 Price distribution for Fragrance and Makeup product



Makeup has the highest price range between 0 and 30 Dollars and mostly of 20 dollars.

The fragrance has the highest price range between 50 and 80 Dollars and mostly around 70 dollars price range.

4.3.7 Top 20 products category having most number of loves



	category	love
0	Liquid Lipstick	179400.000000
1	Contour	83925.000000
2	Eyebrow	71689.473684
3	Eyeshadow	66185.714286
4	Lip Gloss	61055.555556
5	Sponges & Applicators	52291.666667
6	Bronzer	50555.875000
7	Eye Palettes	49660.000000
8	Lipstick	48469.444444
9	Highlighter	45259.888889

This displays the top 10 products with the most likes on the Sephora website from all brands on the website. Lipstick has the highest number of likes of 179400, second Contour product with 83925 likes, and third eyebrow products with 71689 likes.

5.1 CONCLUSION

- After applying Wilcoxon Text it proves that there is a difference between the price and valued price of products on the Sephora website.
  The actual initial value price of the product varies from the price at which the product is sold on the website with seasonal offers and discounts offered.

- We concluded that the number of reviews received on the product makes a difference in rating.
  The product with a high number of reviews and more ratings on them.

- It is verified that there is no association between online sales & exclusive sales of products by brand on the Sephora website.
  Online sales and exclusive sales only on the Sephora website are independent of each other and have no association between them.

- limited edition products are launched less on Sephora than other products.
  The limited edition products are expensive compared to other products of the name brand.

- The marketing flag used for marketing affects the brand name on the Sephora website.
  The products get more exposure when marketing flags are used on them which makes a difference in sales of the product.

- Brands don’t make many limited edition products as they don't sell much compared to other products of the same brand

- Bumble bumble has most products to sell on the Sephora website.

- As described by Sephora, one of the wanna-be retailers, the Sephora website has generally a good or high number of ratings of products sold by its customer.

- Liquid Lipstick is the product category that has the most likes recorded.

- There are more Fragrances and hair styling products offered from all brands and any other kinds of products.
- Most products on the Sephora website are directly launched on the website without any particular marketing as the brands are well-known and trusted.

- Sephora has products mainly ranging from 0 to 100 Dollars.

- The products from different brands are chiefly sold online only then in the Sephora store. We can say the online sales of products are higher than any other method.

6.1 REFERENCES

https://www.kaggle.com/datasets/raghadalharbi/all-products-available-on-sephora-website?resource=download

https://medium.com/aiguys/r-programming-hypothesis-testing-d584f1231836

https://data-flair.training/blogs/hypothesis-testing-in-r/in-r/